Balanced Low-Rank Adaptation:
Removing Parameter Invariance to Accelerate Convergence

Valérie Castin[1] Kimia Nadjahi[1] Pierre Ablin[2] Gabriel Peyré[1]

[1]Ecole Normale Supérieure PSL CNRS

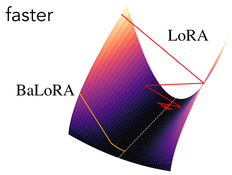
[2]Apple

SCAN ME

BALORA IN A NUTSHELL

We propose a theoretically-grounded variant of LoRA that converges faster

- The LoRA loss $f(A,B)$ has a continuous manifold of optimizers
- **Balanced** minimizers $A^\top A = BB^\top$ are **optimally conditioned**
- When training LoRA, **projecting** the iterates to a **balanced** manifold **accelerates convergence**



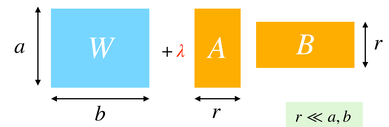
LOW-RANK ADAPTATION (LoRA) [Hu et al., 2022]

Parameter-efficient method to adapt a pretrained model to a new dataset

❄ freeze pretrained weights W

🔥 train low-rank update to W as

$$\underbrace{W}_{\text{frozen}} + \lambda \underbrace{AB}_{\substack{\text{trained}\\ \text{low rank}}} \quad \text{with stepsize } \gamma$$



Training: optimize $\min_{A_1,B_1,\dots,A_L,B_L} \ell(W_1 + A_1B_1, \dots, W_L + A_LB_L)$ with Adam

Loss with one adapter: $f(A,B) = \ell(W + AB)$

✓ reduced memory cost when training

✓ allows to ship the fine-tuned model easily

CONDITIONING OF MINIMIZERS AND ASYMPTOTIC CONVERGENCE RATE

LoRA is overparameterized: $f(AR, R^{-1}B) = f(A,B)$ for $R \in GL(r)$

→ r^2-dimensional manifold of minimizers

Define the conditioning of (A,B) as $\kappa = L/\mu$ with $\begin{cases} L := \lambda_{\max}(\nabla^2 f(A,B)) \\ \mu := \lambda_{\min \neq 0}(\nabla^2 f(A,B)) \end{cases}$

✓ well-conditioned

✗ ill-conditioned

κ controls the **asymptotic CV rate** for GD and sign-GD (Adam without momentum)

$$\theta_t := (A_t, B_t)$$

$$\limsup_{t\to+\infty} \frac{f(\theta_{t+1}) - f(\theta^*)}{f(\theta_t) - f(\theta^*)} \leq \begin{cases} \left(\frac{\kappa-1}{\kappa+1}\right)^2 & \text{if } \theta_{t+1} = \theta_t - \gamma \nabla f(\theta_t) \quad \text{with } \gamma = 2/(L+\mu) \\ 1 - \frac{1}{r(a+b)\kappa} & \text{if } \theta_{t+1} = \theta_t - \gamma \|\nabla f(\theta_t)\|_1 s_t \quad \text{with } \gamma \text{ well-chosen} \end{cases}$$

Q1: Are some minimizers better conditioned than others?

Q2: How to steer the training dynamics towards an optimally-conditioned minimizer?

THEORETICAL ANALYSIS OF THE CONDITIONING

Model 1: $f(A,B) = \frac{1}{2}\|Z - AB\|_F^2$ (1-layer linear NN)

Proposition: for any minimizer (A,B),

$$\frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2)} \leq \kappa(\nabla^2 f(A,B)) \leq \frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2) - \sigma_{r+1}(Z)}$$

Model 2: $f(A,B) = \frac{1}{2}\|h(AB) - Z\|_F^2$ (h deep non-linear model)

Proposition: if $h(AB) = Z$ (interpolating regime), then for any minimizer (A,B),

$$\kappa(\nabla^2 f(A,B)) \leq \underbrace{\kappa(J_h(AB)^\top J_h(AB))^{1/2}}_{\text{additional factor}} \underbrace{\frac{\sigma_1(A)^2 + \sigma_1(B)^2}{\min(\sigma_r(A)^2, \sigma_r(B)^2)}}_{\text{bound for linear case}}$$

THEOREM: the loss is optimally conditioned if (A,B) is **balanced**, i.e. $A^\top A = BB^\top$.

BALANCED LOW-RANK ADAPTATION (BaLoRA)

After each optimizer step, enforce balancing of (A,B) while preserving $f(A,B)$

$$\begin{cases} \tilde{A}_{t+1} = A_t - \gamma \nabla_A f(A_t, B_t) \\ \tilde{B}_{t+1} = B_t - \gamma \nabla_B f(A_t, B_t) \\ (A_{t+1}, B_{t+1}) = P(\tilde{A}_{t+1}, \tilde{B}_{t+1}) \end{cases}$$

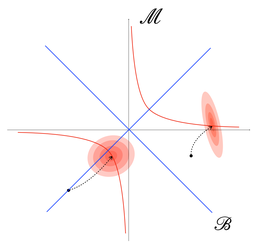
balanced / unbalanced

« Projection » step:

- compute polar decomposition $A = R_A S_A$, $B = S_B R_B$
- compute SVD $S_A S_B = USV^\top$
- apply $P(A,B) = (R_A U S^{1/2}, S^{1/2} V^\top R_B)$

✓ negligible computational overhead

✓ equivalent to a Riemannian GD on the product $X = AB$

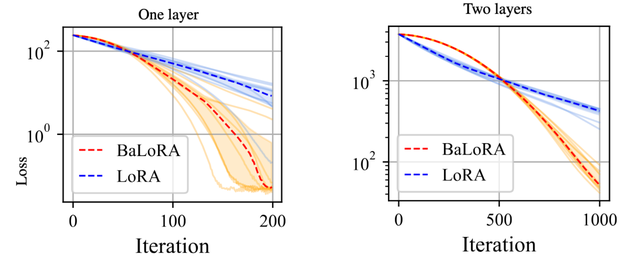


The balanced condition appears in:

- GD and gradient flow for multi-layer NNs [Du et al., 2018; Nguegnang et al. 2024; Marcotte et al., 2023]
- Matrix factorization [Ye & Du, 2021; Ghosh et al., 2025]
- PiSSA [Meng et al., 2024], OLoRA [Büyükakyüz, 2024]

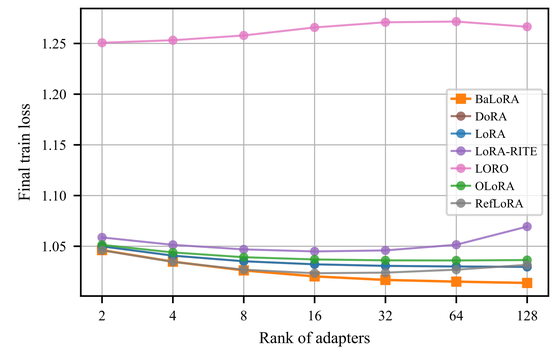
EXPERIMENTS WITH SYNTHETIC DATA

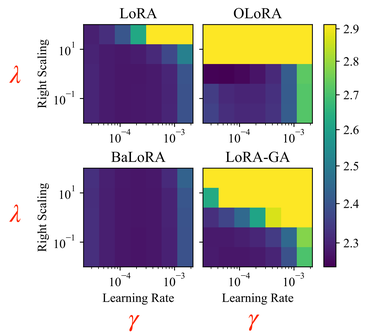
$f(A,B) = \frac{1}{2}\|Z - AB\|_F^2$ $\quad$ $f((A_1,B_1),(A_2,B_2)) = \frac{1}{2}\|W^* - (W_{frozen,1} + A_1B_1)(W_{frozen,2} + A_2B_2)\|_F^2$



EXPERIMENTS WITH LLMS

Rank sweep with Qwen-3.2-3B on DeepMind-Mathematics:





- Balanced methods outperform the others
- BaLoRA has an edge for larger ranks
- BaLoRA works with a wide range of lr and scalings

https://arxiv.org/abs/2605.31484

vcastin.github.io valerie.castin@ens.fr